

基于SOM神经网络的高校图书馆个性化需求挖掘系统研究

刘爱琴¹, 李永清²

(1.山西大学经济与管理学院, 太原 030006; 2.中国石油大学[华东]经济与管理学院, 青岛 266555)

摘要: 借助SOM神经网络聚类算法无参数、精准度高的特点, 本文对山西大学图书馆的Web访问行为进行聚类和优化分析。将聚类行为分为指数函数粗调整和线性函数微调整两个优化阶段, 有效提升聚类速率和聚类效果。基于对用户分析结果的输出, 将用户个人特征信息、用户行为数据以及文献数据库进行筛选整合, 形成可靠性和可用性更高的关联数据集, 并结合语义检索和属性值匹配等技术建构用户个性化服务推荐系统, 进行有效性验证, 实现图书馆内部主题推荐、图书推荐和专家推荐三个子系统的协同。

关键词: SOM神经网络; 聚类分析; 个性化推荐; 关联数据集

中图分类号: G250.73

DOI: 10.3772/j.issn.1673-2286.2017.10.006

1 引言

MOOCs大规模开放课程的出现, 极大地颠覆用户网络学习方式, 图书馆提供个性化服务已经成为现代高效能图书馆建设的重中之重^[1]。朱晓云认为个性化服务就是对用户群进行研究, 根据用户需求从庞大的资源中提取用户真正需要的部分^[2]; 张莉萍则更强调知识整合, 认为图书馆应当着眼用户需要, 积极为其提供完善的知识信息服务^[3]。

现代化高校图书馆拥有大量利于读者学习的纸质和电子资源, 根据用户需求实现资源的个性化推荐, 该功能的实现能有效促进高校图书馆服务质量和高校师生教研质量的综合进步。通过获取用户的Web访问数据对用户行为进行聚类, 是完成个性化推荐的首要步骤。

山西大学图书馆记录了大量学生和老师的访问数据, 其中包含用户IP地址、上网时间、访问时长以及访问资源内容等信息^[4]。本文借助SOM神经网络对读者Web访问行为聚类, 实现对读者有针对性的图书馆资源个性化服务推荐, 从而实现图书馆内部主题推荐、图书推荐和专家推荐三个子系统的协同。

2 神经网络聚类体系构建

2.1 SOM神经网络聚类

聚类指基于一定特征的相似程度分组, 实现组内相似度最大, 组间相似度最小的过程^[5], 常用于信息归纳分析和纠错处理。聚类方法在数据挖掘过程中需要满足能够处理分析大规模数据、有效处理高维数据、对异常数据值不敏感、结果与数据输入次序无关以及结果可行等目标要求^[6]。

BIRCH算法是一种增量性分步聚类方法, 只适用于数据分布呈现凸形等部分特殊情况。DBSCAN算法虽然能够自动处理异常数据, 但是必须建立和依据K-dist图对数据的整合进行处理, 且从时间和存储空间上看, 该方法难以适用大规模数据处理。K-means算法具有很强的伸缩性, 广泛应用于处理类间距离较大的数据, 无法对未知数据进行很好的预测处理。K值的选取对数据处理效果有至关重要的作用, 且基于不同的场景有很大差异, 对于未知数据难以进行准确的K值初始设置, 随着输入数据维度的提升和数据量的增加, 该算法在时间成本上消耗较大。

SOM神经网络算法是通过模拟哺乳动物大脑学习和行为过程开发的一种前馈神经网络聚类算法。相比其他方法, SOM网络基于人工智能神经系统实施聚类, 具有无导师自组织学习、无参数、精准化和稳定性强的特点, 能够将高维数据映射为一维或二维数据, 在聚类时间和聚类效果上比其他方法更好, 主要应用于智慧神经网络构建分析、大数据分析处理和用户个性化服务^[7]。针对高校图书馆, 该方法可用于图书馆文献和用户聚类分析, 即根据用户访问行为的聚类结果推断用户需求, 对用户精准定位, 提供个性化推荐服务。本文将用户个人特征信息、用户行为数据、文献数据库数据进行整合, 形成可靠性和可用性更高的关联数据集, 搭建用户的个性化推荐服务系统。

徐涌等在进行Web用户聚类时采用标准的Kohonen神经网络方法, 但该方法聚类速度慢, 难以应用于规模较大的数据分析^[8]; 段隆振等基于标准的Kohonen神经网络算法, 在数据调整中分为粗调整和微调整, 并采用不同的函数形式来实现整体模型的优化训练^[9]。在解决多兴趣度的问题时, 段隆振等借助隶属度的概念进行解决, 会出现资源重复和超量发掘的情况, 使查全率达到预期效果, 但精准化推荐效果会有折扣^[9]。

本文基于分阶段聚类思想, 选取2016年1月5日—9月30日用户访问的相关数据进行模型优化。通过在粗调整优化阶段采用指数函数, 微调整阶段采用线性函数的形式, 保证聚类收敛的速度和效果。在解决用户多兴趣度问题上, 本文通过构建多个聚类中心, 选取2016年10月1日—2017年1月17日的用户访问数据进行仿真, 仿真结果显示该模型能够很好地选择向用户推荐多样化资源的比例, 提升用户个性化推荐服务的精准度和体验效果。

由于用户行为存在不确定性, 进行用户行为聚类时, 如果根据权值距离进行聚类, 阈值设置太大, 会使聚类结果数据冗余, 难以将不同用户的多样化兴趣比例进行区分; 如果设置阈值太小, 会出现用户行为难以聚类的情况, 不利于个性化推荐服务的开展。所以, 本文选择用户行为与权值向量最小欧式距离进行聚类描述, 以解决上述两个问题。

2.2 数据选取和预处理

实现个性化推荐服务与用户兴趣需求相结合, 需要

筛选大量有价值的信息数据进行SOM网络聚类, 同时构建相应的评价指标体系^[10]。为保证实验结果的可行性和可靠性, 本文根据山西大学图书馆的访问数据选取管理学和企业管理学两个主题的部分研究热点为研究对象进行筛选、聚类分析和数据预测。选择的相关研究指标如表1所示。

表1 用户访问数据指标体系

一级指标	二级指标	指标序号
用户访问时长	-	x_1
管理学文献用户阅读量	信息管理与信息系统	x_2
	科技与创新管理	x_3
	组织管理与决策	x_4
	决策理论与方法	x_5
	物流管理	x_6
	系统优化评估	x_7
	经济发展管理	x_8
	生产运营管理	x_9
企业管理学文献用户阅读量	企业战略管理	x_{10}
	技术创新组织与管理	x_{11}
	风险与投资管理	x_{12}
	人力资源管理	x_{13}
	市场营销管理	x_{14}
	财务管理	x_{15}

为避免数据错误带来的影响, 本文以周为单位, 选取调查期内访问时长在30—120分钟的用户访问数据进行分析。其中, 总文献阅读数量=下载文献量+在线浏览文献量, 初步筛选显示可用数据信息有3 211条。

2.3 网络模型

基于原始SOM的网络构架, 将网络体系划分为输入和输出两个层次。其中, 输入层包含 m 个输入神经元, 对应一个一维 m 阶序列; 根据Web预处理信息规划, 将输入层向量中 m 所对应的值设置为15, 输出层中所对应 n 的值设置为8, 故输出层共包含64个神经元用于聚类竞争输出。在模型中, 一维输入层和二维输出层的所有神经元间存在纵向全连接, 输入层中各神经元间相互独立, 不存在权连接; 在输出层中, 为更好地对模型训练, 部分神经元间通过横向权连接来实现反馈作用, 以提升模型训练的精准度和速度。

2.4 算法模型

通过自组织特征神经网络算法并结合模糊数学中隶属度相关概念,将整个算法学习过程分为三步。第一步,粗调整学习运算,通过计算来确定输出层中对应获胜的神经元坐标,实现初步聚类;第二步,提升聚类速度,在邻域缩小的过程中采用指数函数作为邻域函数;第三步,微调整学习运算,将获胜神经元进一步集中优化,以获取更精准的聚类关键点。在此过程中,需要对较集中的连接神经元间的权值进行优化调整,以解决聚类精准度低、学习速率相对较慢的问题。在邻域函数权值优化调整中,不断采用线性递减函数,直至最终模型训练优化完成。如图1所示, $x_1, x_2, \dots, x_m$ 表示输入层的 m 维输入向量, $\omega_{11}, \omega_{12}, \dots, \omega_{mn}$ 为 n 维对应输出层全连接权系数向量。

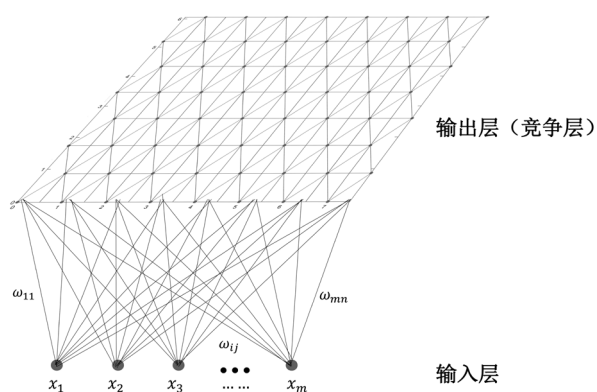


图1 SOM自组织神经网络

2.4.1 获胜神经元的选取

模型的竞争优化过程,实质是通过计算输入向量和全连接权向量间的相似性来进行筛选优化。常见可用性较高的度量输入向量和权向量间相似程度的计算方法为欧式距离法和余弦法^[11]。本文选择输入和权值系数向量相似度最高(即欧式距离值最小)的输出层神经元作为获胜神经元,获胜神经元比其他输出层神经元存在更高程度的优化,并依据相应规则进行权值处理,实现整个模型训练优化效果。

2.4.2 模型优化

整个算法模型优化分为初始模型构建、数据预处理和调整优化三个阶段。当达到初始设置的优化学习次数后,存储并导出全部全连接权系数,模型优化结束。

初始模型构建阶段,整个SOM神经网络结构模型采用MATLAB语言表达,并对输入层与输出层、输出层神经元间的数量和相互关系进行定义。输入层与输出层神经元间的全连接向量通过MATLAB软件随机定义生成,同时保证各权连接向量间线性无关,为后续模型权值优化提供初始环境。

数据预处理阶段,对初始全连接向量和输入向量进行归一化调整处理。归一化处理是一种无量纲化数据处理方式,使各向量间主要表现为相对关系,以保证在向量间进行下一步运算时的科学性和可用性,减少非量纲因素对试验结果的干扰。

调整优化阶段是整个SOM神经网络算法的核心,是决定实际应用效果的关键。高校图书馆资源和使用人员众多,采用标准Kohonen模型中的线性函数优化难以满足时效性要求,本文采用将模型优化过程分为粗调整优化和微调整优化两个阶段的方式,并根据不同阶段的特征进行邻域和学习率调整。

粗调整优化阶段的任务是迅速实现初步聚类,节约优化调整时间。学习函数是粗调整优化阶段主要的操作单元,该阶段采用指数函数作为学习率调整函数,进行500次调整,实现初步输出层神经元聚类。初始阶段设置较大的邻域能保证初始输入神经元有更大的影响效果,提升收敛速度;随着学习率的调整,采用指数衰减法来迭代调整邻域尺寸,以保证聚类的稳定性。

微调整优化阶段主要保证聚类优化效果的可靠性,减少错误和误差对试验结果的影响。此阶段采用线性函数进行学习率调整,进行300次计算。在根据邻域函数进行权向量调整时,借助墨西哥草帽函数模型进行优化调整。

2.5 模型应用

针对模型优化结果,在进行试验应用时,需要遵循以下步骤:第一,向量归一化,即将输入向量和权连接向量进行归一化调整,以保证不同向量间比较时的科学性和可用性;第二,选取输入向量并计算该向量与所有连接权向量间所对应的欧式距离;第三,通过计算比较得出输出层获胜神经元即为该输入向量所属的聚类类别点。

2.6 实验结果

本实验模型优化算法结合标准Kohonen算法优势

进行一定程度改进,在粗调整阶段采用指数函数,微调阶段采用线性函数,并借助墨西哥草帽函数模型调整邻域强度,以提升收敛效果和可靠性。通过对前期搜集数据进行分析优化,形成可对用户行为聚类的代表性神经网络系统。

本文算法在Intel(R) Core(TM) 2/4G RAM/windows 10/MATLAB R2011b进行500次粗调整优化和300次微调优化,再进行样本仿真。为便于对神经网络聚类效果进行可靠性验证,在2016年10月25日—12月30日用户访问数据中随机抽取560条数据进行聚类。由于选择输入和输出神经元的维数限制,最终聚类结果以示意图形式给出,对应的输入向量和激活神经元模型仿真结果如图2所示。结果显示,共有63个神经元被激活,分别代表不同输入向量个体对应的聚类簇,保存到相对应的数据库存储。

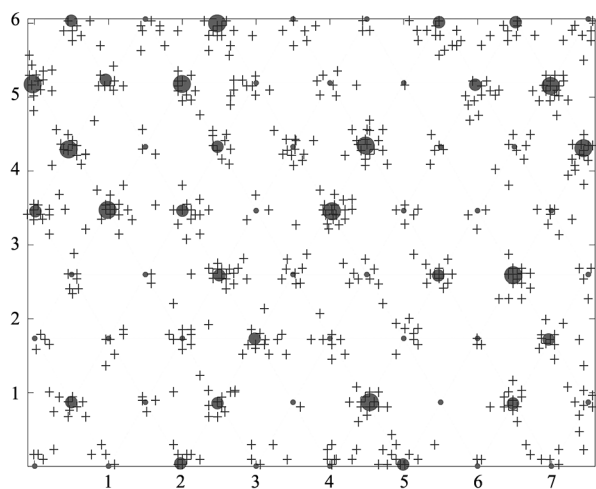


图2 输入向量神经聚类结果

3 个性化推荐服务实现

3.1 实体数据描述

实施图书馆有针对性的个性化推荐服务,仅对Web访问行为聚类还不够,必须将聚类结果与用户个人特征及图书馆文献数据进行对应。

(1) 用户实体特征信息描述。在实验中,山西大学图书馆主要面向的用户为教师和学生,针对不同群体,需要搜集用户的所属群体特征,对用户进行定位。根据个性化推荐服务需求,通过SQL语言选择访问用户实体对应有价值的属性实体进行分析。

(2) 用户Web访问数据聚类结果描述。SOM神经

网络的聚类模型通过粗调整和微调两阶段优化后,15个输入层神经元对应64个输出神经元内部以及两层连接的权向量已经构成聚类系统。由于用户搜索存在不确定性和模糊性,需要通过借助行为聚类分析更好地获知用户需求,并借助EDOAL语言标准对行为聚类结果进行描述。

(3) 图书馆文献检索结果描述。语义检索技术是基于人工智能的一项新兴技术,通过对文献资源数据构建语义数据库,根据用户的需求进行智能化检索。在获得用户的聚类结果后,本文通过搜集此类用户的检索语义和关键词,对用户进行个性化文献筛选和传递^[12]。

3.2 实体链接发现与构建

原始数据包含很多有价值的信息,然而这些数据并不全面,同时包含无价值的噪音数据。为实现个性化服务和推荐的目标,必须对筛选后的用户信息数据集、用户兴趣属性集和文献检索结果描述集整合链接^[13],具体发现过程模型如图3所示。在进行数据实体链接和发现过程中,需要对相关属性和链接进行匹配,根据对应的实体数据选择相应的属性作为原码和外码,最终将多个数据集进行整合,形成一个更完善的数据集合体。为保证数据链接的有效性,要求描述实体信息的相关属性和类不为空。

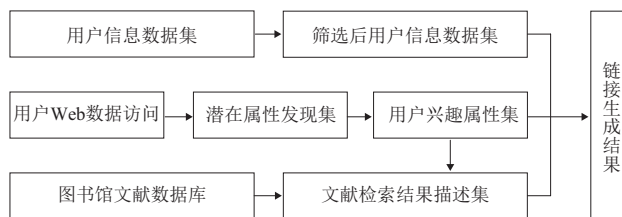


图3 实体链接发现过程示意图

基于此模型,实体链接匹配后,通过SOM神经网络算法对用户实体根据相似度进行自动聚类。神经网络聚类能很好地完成用户间的相似性比较,在进行模型优化时能减少不同属性间的比较时间,提升实体聚类的匹配效率和准确性^[14]。

3.3 个性化推荐服务实现

通过集合以上三种数据,形成完善的综合数据集。该数据集主要包括用户基本信息、用户聚类信息以及用户聚类检索关键词查找的对应文献资源相关信息。

从宏观上看,该数据集合可以展现用户个性化需求;从微观上看,可以使数据集内部的实体属性等信息进行多样化描述。

对用户提供有针对性的个性化推荐服务具体流程包括:(1)选择要进行个性化服务的实体类目,获取对应有价值属性的信息集合;(2)获取该属性实体对应的

聚类信息,如果没有获得有效的用户聚类信息,需要再一次将相关数据通过神经网络系统聚类分析得出最终结果;(3)对获得的信息进行整理,依据检索关键词与数据文献库进行匹配对比,将结果以直观形式展现给用户(见图4)。

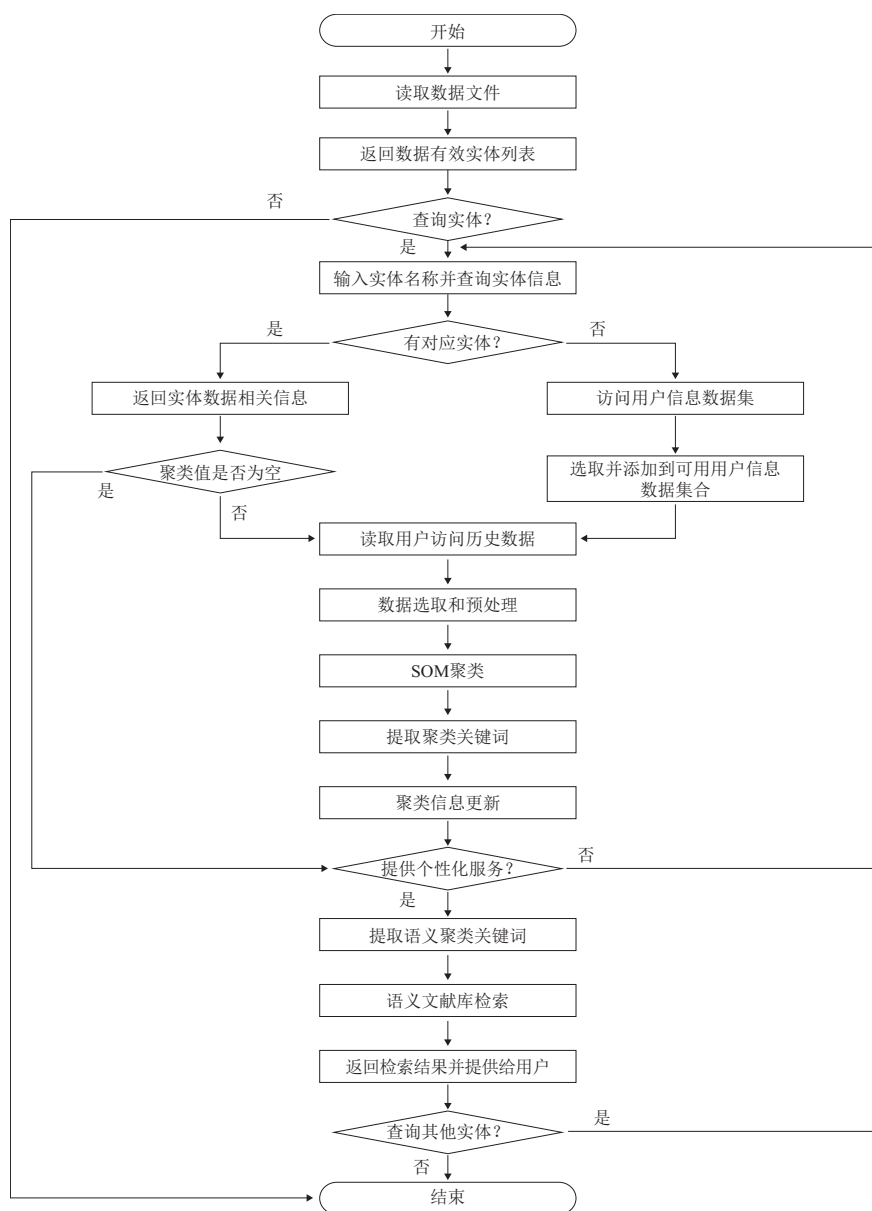


图4 服务个性化推荐过程

3.4 模型分析

在针对性的个性化推荐服务系统建设中,SOM模型的构建对整体可用性起关键作用。本模型在优化过程中采取5种方案。(1)选取时长在30—120分钟的用户访

问数据进行分析,结果显示不但有效规避特殊数据对模型优化的影响,而且提升了整个模型的精准性。(2)由于用户个体的差异较大,通过提供多个竞争层神经元,分别进行500次粗调整优化、300次微调优化,结果显示该模型能够显著提升用户聚类行为的可靠性。(3)通

过对粗调整和微调整采用不同的学习函数,在提升模型优化速率的同时保证权值向量优化的有效性。(4)通过对不同学科构建不同的模型优化子系统,有效地避免在权值优化时向量太多、运算时间太长和不同学科间权值向量互相影响带来的总体优化结果可用性弱的问题。

(5)通过弱化阈值的使用,可对用户的兴趣点推荐进行比例规划,避免资源推荐冗余,提升总体服务质量。

此外,本模型需要结合语义搜索才能发挥其最优作用效果。SILK语言进行语义搜索中的属性值匹配计算,通过RDF来链接和描述实体间的相互关系,并将检索结果输出。本模型通过弱化阈值概念提升查准率,但也降低了查全率,为保证查准率和查全率间的协调性,在实际优化中需要对学习速率、邻域调整函数中的相关参数进行调整。

为进一步提高系统推荐服务的精准性,可对模型进行修正。如在输入数据选取中,进一步修正选择用户文献阅读量的计算方法;在模型优化中,对用户聚类信息保存时长和文献推荐引入时效性概念;定期搜集、整合和筛选用户访问数据,进行个性化反馈推荐。其中,在模型优化阶段,基于现有推荐资源采纳率设定采纳阈值阶梯,并根据用户行为进行系统推荐优化调整。当采纳率值高于最高采纳阈值时,系统自动存储相关数据至一个更新周期,以备后续个性化推荐和为其他用户数据分析作参考;当采纳率值处于最低阈值和最高阈值之间时,针对用户行为采用即时存储、即时优化修正的方式,对用户需求进行实时更新,以提升用户检索结果满意度;针对用户个性化需求偏转和偏移行为,当用户个性化需求低于最低阈值时,系统释放所有存储数据,重新对用户行为进行SOM神经网络聚类分析。

3.5 推荐系统的有效性验证

在图书馆个性化服务推荐中,将推荐方案细分为主题推荐、专家推荐和图书推荐三个不同的子系统。这三种推荐依托于用户兴趣点,分别提供热点资源和专题文献等一系列服务,基于用户不同类型的需要提供有差异的精准推荐。

选取2016年10月1日—2017年1月17日的用户访问数据进行分析,在系统仿真期间,每天对用户的访问数据进行分析并实时更新,通过用户使用过程中的点击量计算系统推荐量数据,计算系统的推荐吻合度。其中,推荐吻合度(%)=用户采纳文献数÷推荐系统推

荐文献数。

从图5可见,在初始阶段三种推荐数据都会有较大波动,不同的系统发展情况有较大差异,随着系统数据的完善和计算更新,整体系统的稳定性逐渐提升,推荐吻合度也呈上升趋势,于第93天后三个子系统的推荐吻合度趋于稳定,实现三个数据子系统间的有效协同。

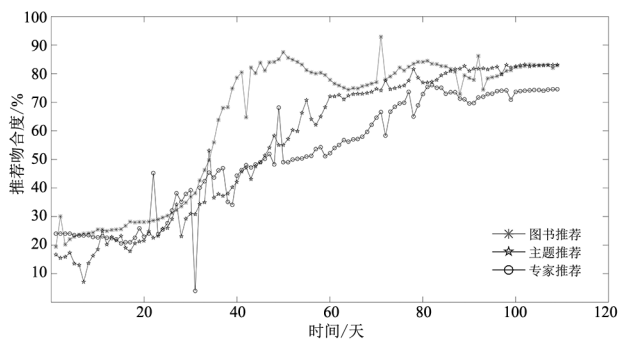


图5 服务个性化推荐系统间的协同

为更好证明系统可用性,本文将新推荐算法与旧检索算法用户采纳率进行对比。选取2015年11月16日—2016年1月4日经济管理类用户检索采纳数据,结果显示,平均用户采纳率为47.53%,用户采纳率主要集中在35%—55%(共33天),其中超过70%的仅为两天,采纳率分别为73.7%和79.91%。相关数据充分证明在有约束情况下系统算法的可靠性。

4 总结

SOM神经网络在数据挖掘中应用广泛,具有无参数、精准度高和客观性强的特点。本文基于原始SOM神经网络算法,对聚类分析过程进行粗调整和微调整两个优化阶段,在保证聚类效果的同时提升聚类速率,并基于修正后的神经网络算法,规划整体个性化推荐服务流程。在个性化服务推荐中,基于用户信息数据集、用户Web访问数据和图书馆文献数据库相关数据资源进行筛选整合,形成关联数据集并进行存储,结合不同用户的偏好和Web访问行为数据,为用户提供基于集群并针对个体的个性化资源推荐。借助现有山西大学图书馆数据,对整体系统有效性进行验证,得出新系统体系优于原始系统体系的结果。

高校图书馆对现代社会发展具有重要作用,优质有效资源的提供对整个高校学科建设和地区经济发展都有非常积极的意义。神经网络算法在现代数据挖掘中应用广泛,相信在现代高校图书馆构建中能够发挥

越来越大的作用。

参考文献

- [1] 刘恩涛,李国俊,邱小花,等.MOOCs对高校图书馆的影响研究[J].图书馆杂志,2014,33(2):67-71.
- [2] 朱晓云.Web数据挖掘与个性化信息服务中用户研究[J].情报技术,2004(2):34-35.
- [3] 张莉萍.论图书馆个性化服务中用户信息的Web数据挖掘[J].情报资料工作,2007(2):101-104.
- [4] 欧阳烽.Web数据挖掘与高校数字图书馆的个性化服务[J].现代情报,2008,28(1):103-104.
- [5] 项冰冰,钱光超.聚类算法研究综述[J].电脑知识与技术(学术交流),2007,2(12):30-31.
- [6] THAMARAISELVI G,KALIAMMAL A.Datamining:concepts and techniques[M],[S.1.]:Morgan Kaufmann Publishers Inc.,2006.
- [7] 张德丰.MATLAB神经网络应用技术[M].北京:机械工业出版社,2012.
- [8] 徐涌,陈恩红,王熙法.基于神经网络的Web用户进行聚类分析[J].小型微型计算机系统,2001,22(6):700-702.
- [9] 段隆振,朱敏,王靓明.基于双Kohonen神经网络的Web用户访问模式挖掘算法[J].计算机工程与科学,2009,31(9):95-98.
- [10] 李赛,邹丽华.神经网络在聚类分析中的运用[J].经济视野,2016(15):238-239.
- [11] 柳胜国.数字图书馆个性化服务与Web日志挖掘数据预处理技术[J].现代情报,2007,27(7):65-67.
- [12] 黄敏,赖茂生.语义检索研究综述[J].图书情报工作,2008,52(6):63-66.
- [13] HIENERT D,ZAPILKO B,SCHAER P,et al.Vizgr:linking data in visualizations[M],[S.1.]:Springer Berlin Heidelberg,2011.
- [14] 高劲松,周习曼,梁艳琪.面向关联数据的实体链接发现方法研究[J].中国图书馆学报,2016,42(6):85-101.

作者简介

刘爱琴,女,1974年生,博士,讲师,硕士生导师,研究方向:信息服务,E-mail: km_aql@sina.com.cn。
李永清,男,1995年生,硕士研究生,研究方向:管理系统优化与战略管理。

Research on Personalized Demand Mining System of University Library Based on SOM

LIU AiQin, LI YongQing

(1.School of Economics and Management, Shanxi University, Taiyuan 030006, China;
2.School of Economics and Management, China University of Petroleum, Qingdao 266555, China)

Abstract: According to the characteristics of high precision and no parameter of the SOM neural network clustering algorithm, the paper, taking the web access behaviors of users in Shanxi University Library as an example, carried on optimized cluster analysis. The progress of clustering behavior could be divided into two stages, the rough adjustment training and the micro-adjustment training, which could improve the clustering rate and effect. Based on the output of analysis results, screening and integrating the user's personal characteristic information, users' behavior data and literature database, to linked data set reliable and available highly. And combining with the semantic retrieval and attributing matching technology, the user personalized service recommendation system was formed and verified effective. It realized the coordination among internal subjects recommending, books recommending and experts recommending.

Keywords: SOM Neural Network; Cluster Analysis; Personalized Recommendation; Linked Data Sets

(收稿日期: 2017-08-26)

《数字图书馆论坛》在2016年度 复印报刊资料转载指数排名中喜获佳绩

由中国人民大学人文社会科学学术成果评价研究中心联合书报资料中心研制的2016年度复印报刊资料转载指数排名于2017年3月28日正式发布。

在“图书馆、情报与档案管理学科期刊”全文转载排名中,《数字图书馆论坛》转载率位列第15名,综合指数位列第20名。

该排名根据人大复印报刊资料近100种学术系列期刊在2015年度转载的学术论文数据,从转载量、转载率、综合指数三个维度对中国人文社科期刊和教学科研机构进行统计形成。